# Instructions

The dataset ‘Cereals.csv’ includes nutritional information, store display, and consumer ratings for 77 breakfast cereals. We are going to apply hierarchical clustering to cluster these cereals based on their nutrition facts.

1. Make sure your Cereals.csv file is in the same folder as your Cereals.R file. Change the working directory use *Session -> Set Working Directory -> To Source File Location* from the Menu Bar.
2. Use read.csv command to import the dataset “Cereals.csv”.
3. Delete the records with missing data. (6 points)
4. Drop the first three non-numerical variables, the three variables about store display: “shelf”, “weight”, “cups” and “customer rating”. (6 points)
5. Normalize the variables using sapply(). (6 points)
6. Apply hierarchical clustering to the data using Euclidean distance and single linkage and plot the tree. (8 points)
7. Apply hierarchical clustering to the data using Euclidean distance and complete linkage. and plot the tree. (8 points)
8. We will use the dendrogram of complete linkage to get the clusters. If we cut at distance 7, how many clusters do we get? (6 points)
9. Get the centroids of each cluster. Which cluster has the lowest calories? Which cluster has the lowest rating? Which cluster would you deem the healthiest in terms of fat, sugar, and sodium levels? (10 points)